Shoulder disability questionnaire: design and responiveness of a functional status measure.

van der Heijden, G.J.M.G.; Leffers, P.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2000

DOI (link to publisher)
10.1016/S0895-4356(99)00078-5

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 53 (2000) 29–38

Journal of Clinical Epidemiology

Shoulder disability questionnaire design and responsiveness of a functional status measure

Geert J. M. G. van der Heijden[a,b,*], Pieter Leffers[b,c], Lex M. Bouter[b,d]

[a]*Institute for Rehabilitation Research, P.O. Box 192, 6430 AD, Hoensbroek, The Netherlands*
[b]*Netherlands School of Primary Care Research, The Netherlands*
[c]*Department of Epidemiology, University of Maastricht, Maastricht, The Netherlands*
[d]*Institute for Research in Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands*

Received 12 November 1998; accepted 2 April 1999

Abstract

The Shoulder Disability Questionnaire (SDQ) is a measure covering 16 items designed to evaluate functional status limitation in patients with shoulder disorders.The responsiveness of the SDQ was evaluated for 180 patients with soft tissue shoulder disorders, without underlying systemic disorders. These patients participated in a randomized placebo-controlled trial, in which ultrasound and electrotherapy appeared to be ineffective as adjuvants to standardized exercise therapy. At baseline and at 6-week follow-up, patients completed the SDQ and rated severity of shoulder pain and their chief complaint, while a research physiotherapist rated severity of symptoms and restriction of mobility. At the 6-week follow-up, patients also rated overall change since baseline. According to the calibrated responsiveness ratio (CRR) and the area under the receiver-operator characteristic curve (AUC) the SDQ discriminates accurately between self-rated clinically stable and improved subjects. The presented results suggest that the SDQ is as responsive as the compared outcome measures, and therefore is ready for use in clinical trials.

Keywords: Shoulder; Physiotherapy; Health status assessment

1. Introduction

Shoulder disorders may limit functional status, impairing the ability of patients to perform functional activities in daily life in a normal manner. Improvement of functional status is an important goal in the treatment of patients with shoulder disorders, and may reduce restriction of socially defined tasks and roles. Therefore, functional status measures are essential in the evaluation of treatment outcome. Yet, in clinical trials on shoulder disorders, little attention is given to functional status measurement [1–3]. In our opinion, it is fundamental that functional status measures are brief, take a short time to complete, and allow self-assessment and completion by patients. Besides covering all crucial elements of functional status, the components of functional status measures must focus on the actual execution of activities that are important in daily life, rather than on the desire to perform them or the perceived possibility to perform them. In addition, functional status measures must be responsive. This means that they must be able to detect clinically relevant changes—changes which are not due to measurement error and biological variability. A summary score should facilitate statistical processing and simple, direct interpretation [4].

Most of the available assessment instruments for shoulder disorders are described in comprehensive reviews [5–7]. These instruments focus predominantly on assessment of symptoms—such as pain, restricted mobility, and decreased muscle power—that are considered to be correlates of functional status limitation. When an evaluation of functional status limitation is included, it typically addresses the clinician's appraisal of the possibility to perform daily activities, instead of the patient's self-report of their actual execution. Most of the available instruments are developed in relation to specific interventions for specific conditions, and little is known about their applicability and responsiveness in other situations.

This paper describes the design of the Shoulder Disability Questionnaire (SDQ), a measure developed to evaluate

* Corresponding author. Tel: +31-45-523-7638; Fax: +31-45-523-1550.
E-mail address: g.vanderheijden@irv.nl (G.J.M.G. van der Heijden)

0895-4356/00/$ – see front matter
PII: S0895-4356(99)00078-5

functional status limitation through self-assessment by patients with soft tissue shoulder disorders who participate in a randomized placebo-controlled trial on ultrasound and electrotherapy as adjuvants to exercise therapy. Since many outcome measures in our clinical trial were already patient-preferred measures, we wanted to include a measure that was preferred by the physical therapists. Almost simultaneously with the SDQ, several functional status questionnaires for shoulder disorders [8–10] have been developed, which all appear to focus on the patient's perspective for evaluation of functional status limitation due to shoulder disorders.

This paper also documents the responsiveness of the SDQ on the basis of secondary analysis of data from this trial, in which treatments turned out to be ineffective for soft tissue shoulder disorders [11].

2. Questionnaire design

2.1. Item generation

A pool of 60 candidate items was selected from routine history of patients with shoulder disorders in physiotherapy and by reviewing various functional status measures. Suitable items had to refer to activities in daily life involving the upper extremities. In order to focus items on the shoulder, we attached a subclause to each item that explicitly referred to the shoulder. Items were phrased so that they addressed the actual execution of a particular activity, rather than focusing on the desire to perform it or the perceived possibility to perform it. Items which included positive, negative, or ambiguous expressions were rephrased neutrally. Likewise, items with limited applicability were rephrased in more general wording. For example, *"My shoulder hurts when I put a wallet in my back pocket"* was rephrased as *"My shoulder hurts when I reach towards my buttocks"*, while *"My shoulder hurts when I fasten my bra"* was changed into *"My shoulder hurts when I reach towards the back of my chest."*

2.2. Item reduction

We aimed at a questionnaire of approximately 15 items, and decided to reduce the number of items via two surveys. Item reduction surveys were aimed at maximizing the validity of the SDQ according to the judgmental approach, as described by Guyatt and co-workers [4,12]. Since many outcome measures in our clinical trial already corresponded with goals of patients, we wanted to include a patient-completed outcome measure that corresponded with treatment goals of physical therapists.

Therefore, we mailed the list of 60 candidate items to 273 seasoned physiotherapists working in private primary care practices in the south of the Netherlands and to 47 researchers who had published papers on shoulder disorders in Dutch (para)medical journals in 1990–1991. The addressees were asked to select 15 items concerning functional status limitation mentioned most frequently by patients with shoulder disorders. Participants were asked to focus on symptoms and complaints, rather than on attribution of symptoms to an underlying disorder. Next, they had to estimate the impact of each selected item on functional status (Likert scale: 1–5; not at all limiting to extremely limiting). Participants were encouraged to comment on the list of items and to add any items that they felt had been left out. The response to this mailing was 56% for physiotherapists and 60% for researchers. Items were ranked according to the weight-frequency product that was calculated from the returned questionnaires by multiplying the endorsement frequency with the mean item score. We excluded items with a frequency-weight product lower than the median (approximately 200 points; mean item weight = 2.5).

The ranking of the remaining 30 items revealed only minor differences between the 2 groups of respondents. Guided by comments, unclear or confusing items were rephrased. During the second postal survey, the 273 physiotherapists and 47 researchers were asked to select, out of these 30 items, the 15 that they considered most crucial in the evaluation of treatment outcome. Next, participants had to estimate the sensitivity to change of each selected item of functional status limitation (Likert scale: 1–5; not at all sensitive to change to extremely sensitive to change). Again, the respondents were encouraged to comment on the list of items and to add any items that they felt had been left out. The response rate for this mailing was 55% for physiotherapists and 72% for researchers. Items were ranked according to the weight-frequency product calculated from the returned questionnaires through multiplication of the endorsement frequency by the mean sensitivity to change. The median frequency-weight product was approximately 300 points (mean item weight = 2.5). Since the frequency-weight products of items14–20 were very close to the median, the draft SDQ consisted of 20 items. The two groups of respondents showed virtually no differences in the ranking of items. The majority of these 20 items were activities that involved pushing, pulling, reaching, leaning or carrying.

2.3. Finalization for clinical use

Yes-no answer options were used, where *yes* meant that the patient was restricted with respect to the particular activity. To improve accuracy, the recall period in the patient instructions was limited to the previous 24 hours. The answer option *not applicable* (NA) was added to focus patients on the actual execution of activities, rather than on the desire or the perceived possibility to perform them. A NA response meant that the activity of the particular item (e.g., carrying something) had not been performed in the previous 24 hours. The ratio of the number of items with an affirmative answer over the number of applicable items was multiplied by 100. This ratio was used as summary score and ranged from a maximum of 100 (i.e., affirmative answer to all applicable items) to 0 (no functional status limitation).

In order to evaluate the applicability of the items and the appropriateness of the answer option, 12 patients with shoulder disorders treated by physical therapist completed the draft SDQ. They were asked to comment on the format and item wording, and add any items that they felt were missing. Their comments resulted in refinements in the wording of five items. Eight items addressing similar upper extremity activities were combined: writing and typing, opening and closing a door, gripping a steering wheel or bike handlebars, and putting on a coat or a sweater. A preliminary English translation of the SDQ can be found in the Appendix.

3. Methods

3.1. Patients

The responsiveness of the SDQ was evaluated in 180 patients. Between 1 May 1992 and 1 November 1994 these patients were enrolled in a randomized placebo-controlled trial for patients with soft tissue shoulder disorders in which ultrasound (US) and electrotherapy (ET) had been ineffective [11]. Eligible patients had either or both (1) pain in the deltoid region during glenohumeral movement, and (2) restricted passive range of glenohumeral motion, while it was very likely that they had either or both (3) a localized soft tissue lesion, and (4) involvement of the sympathetic nerve system. Excluded were patients who had a stroke, polyneuropathy, multiple sclerosis, rheumatoid arthritis, polymyalgia, ankylosing spondylitis, malignancy, hemophilia, prior fractures, or prior surgery, along with those having motor or sensory deficits, or wounds or skin defects in the shoulder, upper limb, neck, or thorax. Furthermore, patients who were considered to have major shoulder hypermobility, complete rotator cuff tears, glenohumeral joint inflammation, or referred pain from the neck or from internal organs in the shoulder were excluded, as well as those who had already received ET or US during the current episode, were completely or nearly completely free of symptoms, or those who indicated reluctance to adhere to the allocated treatment or to complete follow-up.

Additionally, to increase the efficiency of the clinical trial, patients at the positive end of the prognostic spectrum were excluded because it was unlikely that ET and US could hasten their recovery. These were patients with very large improvement, who at the same time fulfilled three or four of the following putative indicators of a favorable prognosis: (1) only dominant side impaired, (2) first episode ever, (3) no pain radiating below the elbow, or (4) no co-existent cervical or elbow disorder. For the same reason, patients at the negative end of the prognostic spectrum were excluded because it was unlikely that they would benefit from any treatment. These were patients without any improvement, who at the same time fulfilled three or four of the following putative indicators of a poor prognosis: (1) non-dominant side or bilateral impaired, (2) prior episodes, (3) pain radiating below the elbow, or (4) co-existent cervical or elbow disorder. All consenting patients received exercise therapy, while according to randomization, some patients received adjuvant physical modalities for only one shoulder. In subjects with bilateral shoulder problems, the shoulder of the dominant arm was treated.

3.2. Data collection

In addition to completion of the SDQ at the end of the qualification period and 6 weeks later, patients gave visual analogue scale (VAS) severity ratings for (1) the chief complaint in the preceding week, and (2) shoulder pain in the preceding week. Simultaneously, based on a standardized clinical assessment, the research physiotherapist gave VAS severity ratings for (3) symptoms, and (4) mobility restriction. The legend of Table 1 provides a more detailed description of these measures. All ratings ranged from 0 (minimum) to 100 (maximum). At 6 weeks, patients rated overall change since baseline on an 8-point Likert transitional scale. This rating was used as an external criterion for the analysis of responsiveness. Complete recovery, very much, and much improved was referred to as *clinical improvement*; little improved, unchanged, and little worse as *clinical stability*; while much and very much worse was referred to as *clinical deterioration.* Data were analyzed with SPSS for Windows (version 6.1.2).

3.3. Ability to detect change

We evaluated the individual SDQ items for those capable of detecting change. For this purpose we plotted for each item the distribution of change in answer options since baseline, either improvement or deterioration, for all 180 patients. We also explored potential floor and ceiling effects (i.e. the inability of an outcome measure to detect change towards the end of its scale) given the room for change [13]. The presence of a floor effect was studied in subjects with the lower quartile or the best SDQ baseline summary scores, by calculating the proportion of clinically improved subjects with an improvement in SDQ score since baseline. The presence of a ceiling effect was studied in subjects with the upper quartile or the worst SDQ baseline summary scores, by calculating the proportion of clinically deteriorated subjects with a deterioration in SDQ score since baseline.

The ability to detect any change of an outcome measure can be assessed by an effect size statistic (ES) [14–20]. For this, no external criterion is needed. In a single group design, the mean change in score in the population is used as the numerator of the ES, while its denominator is expressed as the associated standard deviation [14–18] or the standard deviation of the baseline score [17]. In a control group design, the difference in mean change scores of the groups is used as the numerator of the ES, while its denominator can be expressed as the (pooled) standard deviation of the baseline scores of the population [14–18]; the standard deviation of the change in the control group [14–18]; or the standard

Table 1.
Clinical characteristics of subjects

	All $n = 180$	Stable $n = 82$	Improved $n = 92$
Age (mean ± SD)	51 ± 13	51 ± 12	51 ± 14
Females (%)	51	49	52
Previous episodes (%)			
None	56	44	65
1–2	27	30	26
>2	18	26	9
Co-existent disorders (%)			
Cervical	86	85	86
Homolateral elbow	68	70	66
Impaired shoulder (%)			
dominant	50	50	51
right	59	56	62
Duration prior to intake (%)			
0 weeks–3 months	38	35	42
3 months–6 months	27	30	22
6 months–12 months	19	16	21
>12 months	16	18	15
Radiating pain below elbow (%)	73	67	76
Cause current episode (%)			
Trauma	12	26	39
Unknown	52	56	48
Onset current episode (%)			
Acute	32	26	39
Gradual	68	74	61
Symptoms prior to intake (%)			
Stable	10	5	14
Increased	71	72	68
Decreased	19	23	17
Prognostic grading[a,b]	49 (25,69)	51 (30,73)	43 (25,66)
SDQ[a,c]	73 (63,87)	74 (63,85)	70 (58,87)
Shoulder pain[a,d]	54 (40,70)	59 (40,72)	54 (42,67)
Chief complaint[a,e]	73 (57,85)	74 (66,86)	71 (52,84)
Symptoms[a,b]	47 (30,63)	47 (27,69)	47 (30,62)
Mobility[a,f]	44 (20,68)	42 (21,69)	39 (18,67)

[a]Presented are medians with 25th and 75th percentiles between brackets.

[b]Prognostic grading by the research physiotherapist on a visual analogue scale (VAS; 0/100: best/worst), is based on symptom severity combined with presented clinical characteristics. Rating of symptom severity (VAS; 0/100: best/worst) was based on a standardized assessment, including history; inspection of contour, muscle wasting, and swelling; active and passive evaluation of range of motion and shoulder pain on abduction, flexion, internal and external rotation, extension, and adduction; evaluation of the mobility and active glenohumeral-scapulo-thoracic rhythm; evaluation of joint play and shoulder pain on accessory movements; evaluation of muscle weakness and shoulder pain on isometric muscle testing; and palpatory assessment of shoulder pain and tissue condition.

[c]Shoulder Disability Questionnaire (0/100: best/worst).

[d]Severity rating by the patient for the shoulder pain during the preceding week (VAS; 0/100: best/worst).

[e]Severity rating by the patient for the chief complaint during the preceding week (VAS; 0/100: best/worst). The chief complaint was defined as the major unavoidable painful and/or limited daily activity in which the shoulder is involved.

[f]Severity rating for restriction of active mobility by research physiotherapist (VAS; 0/100: best/worst), based on a standardized assessment of (1) glenohumeral-scapulo-thoracic rhythm; reaching with the index finger towards (2) the heterolaterol scapular angulus inferior and (3) the second thoracic processus spinosus; (4) flexion; and (5) abduction.

deviation of the baseline scores of control group [17]. So far no method for calculating ES has become a standard, since none seems to be superior over the other. In general, the use of a standard deviation of baseline scores in the denominator is expected to result in a larger ES, because it does not include response variance.

Because an external criterion for clinically relevant change was available in our study, we could assess the ability to detect clinically relevant change by a Responsiveness Index (RI) [14,18–20]. Its numerator could be expressed as the mean change in score of clinically changed subjects, the mean change in score of subjects receiving a treatment of known benefit, or the minimal clinically relevant difference [18,19]. Its denominator could be expressed as the standard deviation of change in subjects who are clinically stable, receive no treatment at all, or receive a placebo treatment [18,19]. Again, no RI has become a standard, since none seems to be superior over the other.

3.4. Responsiveness

We defined responsiveness as the ability of an instrument to discriminate between clinically stable and improved subjects. Responsiveness is, above all, a function of the variability in clinically stable subjects [19,20]. In the absence of a gold standard for clinical stability, we consider the average change in score since baseline in self-rated clinically stable subjects to be the best estimate of the true value for stability. We explored the responsiveness of the SDQ and other outcome measures by their distribution of change in score since baseline in self-rated clinically stable and improved subjects (Fig. 1). The smaller the range of these distributions and the smaller their overlap, the larger the responsiveness of an outcome measure will be. Because all outcome measures were used simultaneously in the same study population, we evaluated and compared their responsiveness by two parameters: the calibrated responsiveness ratio (CRR) and the area under the receiver-operator characteristic curve (AUC).

3.4.1. The CRR

We consider the ratio of the mean change in score of clinically changed subjects and the standard deviation of change in clinically stable subjects the most suitable RI for our purpose. We calculated non-parametric CRRs with medians and interquartile ranges (IQRs), because the change in score since baseline of clinically stable and improved subjects for the compared outcome measures had non-Gaussian distributions (Fig. 1). Comparison of responsiveness of different outcome measures by this RI, however, may be complicated by systematic changes in score since baseline in clinically stable subjects. We wanted to calibrate the RI of outcome measures for these changes. Therefore, analogous to the numerator of the ES for a control group design, we subtracted the median change score in clinically stable subjects from the median change score in clinically improved subjects. We denote the resulting ratio as CRR: the differ-

ence between the median change in score since baseline for self-rated clinically improved and stable subjects, over the IQR of this change in clinically stable subjects. There is no uniform threshold for the CRR above which an outcome measure can be called responsive. The larger the CRR, the larger the responsiveness of the outcome measure. If an outcome measure has a CRR smaller than 1, the change in scores since baseline of improved minus stable subjects does not exceed the IQR (i.e., all possible variance) in stable subjects, and we consider the outcome measure not responsive.

Conceptually, the CRR is very similar to a Responsiveness Index (RI). The RI, as proposed by Guyatt, can be used to compare the responsiveness of different outcome measures. The assumption underlying the validity of such a comparison is that the mean changes in clinically stable subjects are equal across the compared instruments. Ideally, the mean changes in clinically stable subjects are equal to zero. For several reasons, for example, regression to the mean, these mean changes are neither similar across compared instruments nor equal to zero. Therefore, the comparison of instruments according to their RI is likely to yield biased results. To adjust for this potential bias, we calibrated the RI by subtracting the mean change in clinically stable subjects from the numerator for each instrument. Because the data for the compared instruments have a non-Gaussian distribution, parametric statistics will result in biased CRRs. Therefore, we followed the suggestion of Kazis [17] to use non-parametric statistics for the calculation of effect sizes. When data have a Gaussian distribution, CRR can also be calculated using parametric statistics.

3.4.2. The AUC

In the context of responsiveness, the ROC curve plots the true-positive proportion (Y-axis) against the false-positive proportion (X-axis) [20,21] of clinically improved subjects with a change in score since baseline equal to or larger than a cut-off point in the score range. The AUC for changes since baseline represents the probability of correct discrimination between pairs of self-rated clinically stable and improved subjects [20,21]. For the construction of the ROC curves, we used steps of 10 points change in score since baseline, starting from maximum deterioration (−100 points) to maximum improvement (+100 points). There is no uniform threshold for the AUC above which an outcome measure is responsive. However, if the AUC is equal to 0.5 (i.e., the true and false positive proportions describe a 45° line of identity throughout the score range), it means that the outcome measure does not discriminate between clinically stable and improved subjects. This indicates that the outcome measure is not responsive. The AUC approaches 1.0 if the ROC curve reaches higher and towards the left in the diagram, indicating that the outcome measure approaches perfect accuracy in discriminating between pairs of clinically stable and improved subjects [20,21]. Under the assumption of equal utility of true and false positive, the point most upper left in the diagram represents the best cut-off point, that is, the point with the optimal tradeoff between true and false positive proportion or the highest likelihood for correct discrimination of stable and improved subjects [20,21]. The use of such a cut-off point may facilitate decisions concerning treatment of patients with shoulder disorders.

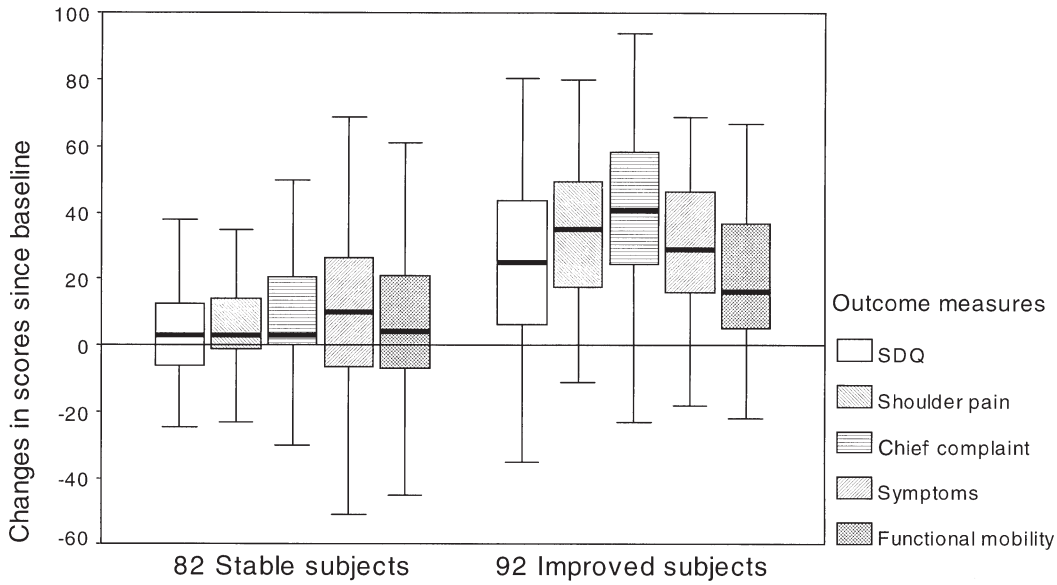

Fig. 1. Box plot for changes in score since baseline of compared outcome measures for stable subjects ($n = 82$) and improved subjects ($n = 92$). Boxes represent the range between 25th and 75th percentiles for the change in score since baseline. Horizontal lines inside boxes represent median change in score since baseline. Tails represent the range between 5th and 25th, and between 75th and 95th percentiles for the change in score since baseline.

4. Results

4.1. Subjects and responses

Table 1 presents the clinical characteristics of the 180 subjects, of whom 92 (51%) were clinically improved and 82 (46%) were clinically stable. All subjects could easily complete the SDQ at baseline and 6 weeks later; completion took between 5–10 minutes. At baseline and 6 weeks later, NA responses were sporadically obtained (modus = 1; range = 1–4). In addition, NA answers were most frequently reported for items 7, 8, 11, and 16. Very few subjects gave an affirmative answer at baseline for items 7 and 16. For all 180 subjects, a change in answer from *yes* at baseline to *no* 6 weeks later or vice versa, was most frequently reported for items 1, 2, 4, 6, and 12–14, while the least change was reported for items 7, 10, and 16 (Fig. 2). Hence, items 7 and 16 yield very little information about change in functional status limitation.

4.2. Floor and ceiling effects

Of the 59 subjects with the best (or lower quartile) SDQ baseline scores (0–63), the change in SDQ scores since baseline exceeded zero in 25 (71%) of the 35 clinically improved subjects and in 8 (33%) of the 24 non-improved subjects. This suggests that the SDQ is also able to distinguish between improved and non-improved subjects when there is relatively little room for improvement of functional status at baseline. Therefore, an important floor effect for the SDQ summary score appears to be unlikely in our study population. Of the 40 subjects with the worst (or upper quartile) SDQ baseline scores (87–100), the SDQ scores since baseline deteriorated in none of the 6 clinically deteriorated subjects (0%) and in 2 of the 34 non-deteriorated subjects (6%). This suggests that the SDQ is not able to distinguish between deteriorated and non-deteriorated subjects when there is relatively little room for deterioration in functional status at baseline. Therefore, a ceiling effect for the SDQ summary score appears to be very likely in our study population.

4.3. Responsiveness

Fig. 1 displays the distributions of change in score since baseline of clinically stable and improved subjects for the compared outcome measures. The median change in SDQ score since baseline for clinically stable subjects is small, while it is large for improved subjects. There is some overlap of the associated IQRs. For shoulder pain and chief complaint there is no overlap of the IQRs around their medians (Fig. 1). The non-parametric CRRs for the SDQ, the chief complaint, and shoulder pain are 1.14, 1.59, and 1.96 respectively. Since the CRRs of symptoms (0.56) and mobility (0.40) are smaller than 1, we consider these outcome measures not responsive.

Fig. 3 displays the ROC curves for the compared outcome measures. We consider them all responsive, since their AUCs all exceed 0.5: SDQ = 0.72; shoulder pain = 0.80; chief complaint = 0.79; symptoms = 0.76; and mobility = 0.67. Under the assumption of equal utility of true and false positives, the SDQ discriminates well between improved and stable subjects when 10% to 60% of the applicable items change in score from yes to no. When 50% of the applicable items change from yes to no, the SDQ has the highest likelihood for correct discrimination between improved and stable subjects. The chief complaint, shoulder pain, and symptoms have the highest likelihood for correct discrimination between improved and stable subjects when VAS scores improve 70, 50, and 40 points of the maximum of 100, whereas functional mobility has poor ability to discriminate between improved and stable subjects.

5. Discussion

We wanted to know whether patients with shoulder disorders perceived change in their ability to perform functional activities in daily life in a normal manner. For this purpose we designed the SDQ, a 16-item measure for functional status limitation in patients with shoulder disorders. The selection of items was based on functional status limitations most frequently reported to, and judged crucial in the evaluation of treatment outcome by relevant health care professionals. The SDQ focuses on how symptoms and complaints of patients with shoulder disorders affect their ability to perform daily activities. In our experience, the SDQ is convenient for patients since it is easy to complete, taking only a little time, while the chosen answer options are easily quantifiable and interpretable, both on item level and as a summary score. We only included subjects with local shoulder disorders. For reasons of efficiency of the design of our clinical trial we excluded 84 subjects because of their favorable prognosis at baseline and 2 for their poor prognostic status at baseline. At the 6-week follow-up, the 2 subjects at the negative end of the prognostic spectrum reported a deterioration of their complaints, while 58% of the other 84 subjects reported very large improvement (including complete recovery)—a recovery rate which was never reached by the trial participants within 12 months. These exclusions clearly resulted in lower estimates of responsiveness of the SDQ and the other instruments. Therefore, we believe that the SDQ is a suitable instrument for assessing functional status limitation, and able to detect clinically relevant change.

5.1. External criterion

There is no gold standard that provides a valid and reliable estimate for clinically relevant change in patients with shoulder disorders. In the absence of such a gold standard, its definition depends on the judgment by either clinician or patient. We consider the patient's self-report to be the best estimate for clinically relevant change. Therefore, we decided to use it as an external criterion in the analysis of responsiveness. Although this external criterion appears to

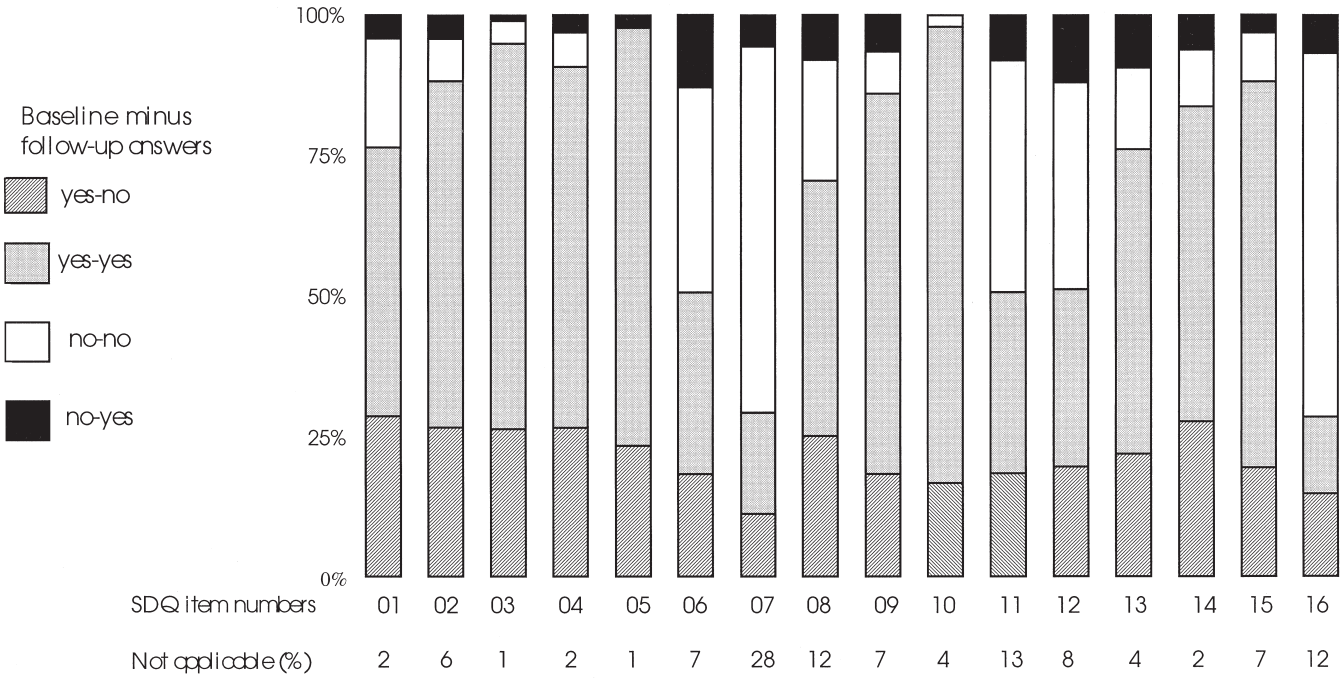

Fig. 2. Bar chart for changes in answers between baseline and follow-up for 16 SDQ items in 180 subjects.

provide a clinically relevant difference, it does not necessarily represent the minimal clinically relevant change. Nevertheless, the SDQ detected relatively small changes in scores since baseline for clinically stable subjects and relatively large changes for clinically improved subjects. The majority of the 180 participants in our study improved, while very few deteriorated. The identification of a ceiling effect in the study population might be explained by the small number of clinically deteriorated subjects. It is obvious that this population cannot be used to demonstrate the ability of the SDQ to pick up deterioration. Therefore, it cannot be concluded from our study that the SDQ is more likely to pick up improvement than deterioration.

We consider the SDQ to be an outcome measure for the

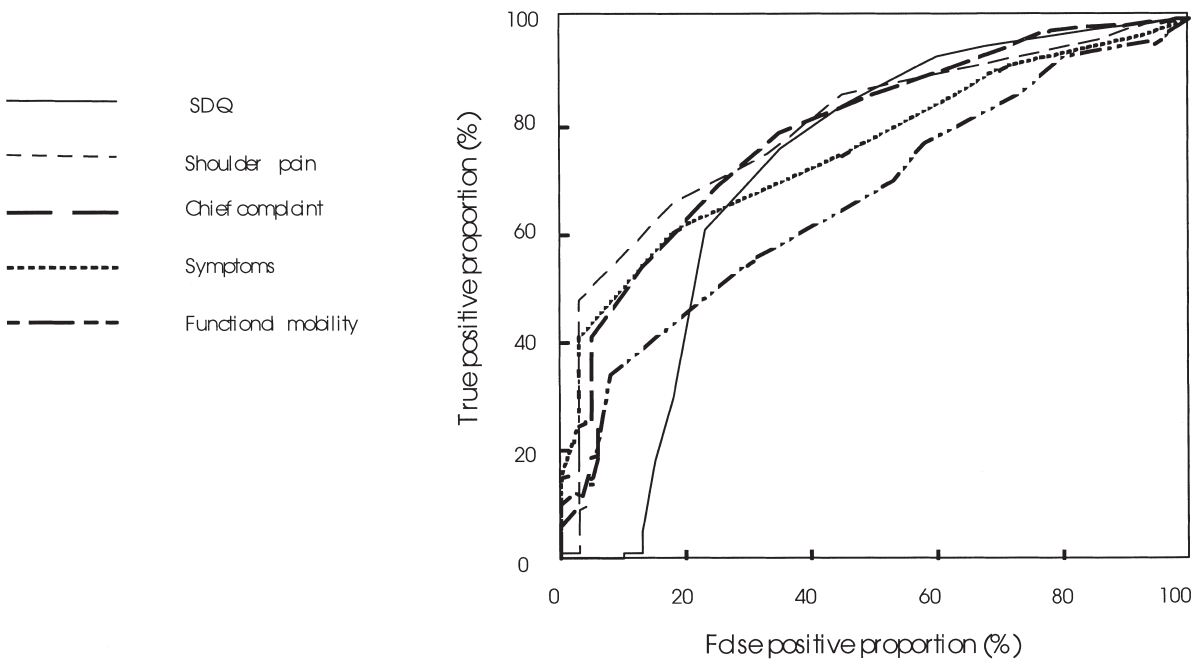

Fig. 3. Receiver operator characteristic (ROC) curves for stable subjects ($n = 82$) and improved subjects ($n = 92$) at multiple cut-off points for changes in score since baseline of compared outcome measures.

severity of functional status limitation in patients with shoulder disorders that is as responsive as any of the compared outcome measures. The highest likelihood of accurate discrimination between clinically stable and improved subjects for the SDQ in this study is established when 50% of the applicable items change in score from yes to no, while the smallest improvement with the highest likelihood of accurate discrimination between stable and improved subjects proved to be a change in score from yes to no for 10% of the applicable items.

Test-retest reliability of the SDQ appears to be sufficient; that is, there is little variation in scores in patients who, according to our external criterion, are considered to be clinically stable. Our study, however, is not designed to demonstrate the discriminative ability of the SDQ for high and low functional status scores. For this, a study with a cross-sectional design would be needed, with sufficient numbers of subjects with high and low functional status scores that are stable over time.

Differences in the nominator due to many NA responses may have implications for the interpretation of SDQ scores (e.g., affirming 10 out of 10 items may not reflect the same functional status limitation as affirming 10 out of 16 items). But it is not clear whether the number of NA responses is related to the severity of the shoulder disorder, or whether a reduction in the number of NA scores reflects improvement. When, for a specific item, a NA response was given at baseline, 6-week measurement, or both, it was impossible to calculate change scores since baseline. An alternative approach to including items with an NA response in the analysis, is to replace NA responses by the previous or following response, its opposite or both. This allows us to perform a sensitivity analysis, in order to estimate the influence of NA responses. In our study, very few NA responses were obtained. A sensitivity analysis did not show major influence of NA responses with respect to the CRR and AUC (data not shown). We did not use a nominal or an ordinal response scale to evaluate why or to what extent certain activities were not carried out. Now, since its applicability and responsiveness is established, it would be worthwhile to see how such changes in scaling might improve the performance of the SDQ.

5.2. Analysis

The relative ranking of the outcome measures according to the CRR and AUC is quite similar. Thus, our results are not the result of expressing responsiveness according to one or the other method. The differences between instruments, however, are more apparent with the CRR than with the AUC. This fact is also reported in other responsiveness studies [14,22,23]. One explanation for the differences in responsiveness is the difference in recall period: 24 hours for the SDQ versus 1 week for the other outcome measures. The most likely explanation for the differences in responsiveness between the compared outcome measures, however, is the difference in the variance of the change in score since baseline in clinically stable subjects. The differences in these variances are likely to be determined by the nature of the compared outcome measures. For example, the chief complaint is an individually-tailored measure for functional status limitation that includes only one activity of daily life. In addition, the compared measures will show different recovery patterns. For example, shoulder pain is known to subside rapidly over a short time in many patients, while mobility restriction resolves only slowly.

5.3. Similar instruments

Almost simultaneously, three functional status measures were developed that are similar to the SDQ. The Disability Questionnaire by Croft et al. [8], includes 22 items with a yes-no answer scale and a 24-hour recall frame. The items of the Disability Questionnaire concern functional activities and movements with the arm. The Shoulder Pain and Disability Index (SPADI), developed by Roach et al. [9], consists of a separate 5-item pain scale and an 8-item disability scale, with the preceding week as the recall frame. In order to make the SPADI suitable for telephone administration, the original visual analogue answer scales have been converted into 0–10 numerical scales [24]. The Shoulder Rating Questionnaire by l'Insalata et al. [10] consists of 19 items with a 5-point ordinal answer scale: 4 relate to pain, 6 to daily activities, 3 to recreational and athletic activities, 5 to work, and 1 to satisfaction. The Shoulder Rating Questionnaire also includes a visual analogue scale for global assessment, as well as an item to indicate the domain of most important improvement. The Shoulder Rating Questionnaire has a recall frame of 1 month. As with the SDQ, these three measures all include items that refer to problems with sleeping, dressing, and functional activities and movements with the arm. The overlap with respect to item content is largest between the SDQ and Croft's Disability Questionnaire. Van der Windt et al. evaluated the responsiveness of the SDQ during a survey of primary care in the Netherlands. She reported AUCs of 0.84 and 0.90 at 3 and 6 months follow-up, respectively [23]. The responsiveness of the SPADI was established in primary care in the United States by Williams et al. [24] and Heald et al. [25] Williams et al. reported an AUC of 0.91 at 3-month follow-up [24] while Heald et al. reported a standardized response mean of 1.38 [25]. For their evaluation of responsiveness, Williams, Van der Windt, and Heald also used patient report of improvement as external criterion.

5.4. Epilogue and recommendations

The results reported in this paper are based on the Dutch version of the SDQ in a population of patients with soft tissue shoulder disorders in primary care physiotherapy. Although they require confirmation by other investigators, these results suggest that the SDQ is at least as responsive as the compared outcome measures, and ready for use in

clinical trials. It is very likely, however, that the outcome of the SDQ and the compared measures will be different in a population in another health care setting or with another cultural background and language. Therefore, a formal translation [26] is indicated before using the preliminary English translation of the SDQ (Appendix).

It would be worthwhile to evaluate the responsiveness of the SDQ and its potential floor and ceiling effects in another population and health care setting, relative to that of other measures for functional status limitation of the shoulder. Furthermore, because it has been shown to improve responsiveness of other disease-specific functional status questionnaires [27,28] reduction of the number of items (e.g., excluding items 7 and 16 because they yield hardly any information), as well as the use of VAS or Likert scales, should be given attention in future studies. Moreover, before the SDQ is used for cross-sectional discriminative purposes, test-retest reliability needs to be evaluated separately.

References

[1] van der Heijden GJMG, van der Windt DAWM, de Winter AF. Physiotherapy for soft-tissue shoulder disorders. A systematic review of randomized clinical trials. Brit Med J 1997;315:25–30.

[2] van der Heijden GJMG, van der Windt DAWM, Kleijnen J, Koes BW, Bouter LM. Steroid injections for shoulder disorders: A systematic review of randomized clinical trials. Brit J General Practice 1996; 46:309–16.

[3] van der Windt DAWM, van der Heijden GJMG, Scholten RJPM, Koes BW, Bouter LM. The efficacy of non-steroidal anti-inflammatory drugs (NSAIDS) for shoulder complaints: A systematic review. J Clin Epidemiol 1995;48:691–704.

[4] Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. Can Med Assoc J 1986;134:889–95.

[5] MacDonald DA. The shoulder and elbow. In: Pynsent PB, Fairbank JCT, Carr A, editor. Outcome Measures in orthopaedics. Oxford: Butterworth-Heinemann, 1993. pp. 144–73.

[6] Beaton DE, Richards RR. Measuring function of the shoulder: A cross-sectional comparison of five questionnaires. J Bone Joint Surg 1996;78(A):882–90.

[7] Stock SR, Cole DC, Tugwell P, Streiner D. Review of applicability of existing functional status measures to study of workers with musculoskeletal disorders of the neck and upper limb. Am J of Industrial Med 1996;29:679–88.

[8] Croft P, Pope D, Zonca M, O'Neill TO, Silman A. Measurement of shoulder related disability: results of a validation study. Ann Rheum Dis 1995;53:525–8.

[9] Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. Arthritis Care Res 1991;4:143–9.

[10] L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MGE. A self-administred questionnaire for assessment of symptoms and function of the shoulder. J Bone Joint Surg 1997;79(A):738–48.

[11] van der Heijden GJMG, Leffers P, Wolters PJMC, Verheijden JJD, van Mameren H, Houben JP, et al. Efficacy of ultrasoundtherapy and electrotherapy for shoulder disorders: Results of a randomized placebo-controlled trial. In: van der Heijden GJMG, editors. Shoulder disorder treatment: Efficacy of ultrasoundtherapy and electrotherapy. [PhD thesis]. Maastricht: Datawyse/University Press Maastricht, 1996. pp. 79–91.

[12] Jaeschke R, Guyatt G. How to develop and validate a new quality of life instrument. In: Spilker B, editor. Quality of Life Measurements in Clinical Trials. New York: Raven Press, 1990. pp. 47–57.

[13] Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: The floor phenomenon. Med Care 1990;28: 1142–52.

[14] Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–78.

[15] Liang MH, Fossel AH, Larson MG. Comparison of five health status instruments for orthopaedic evaluation. Medical Care 1990;28: 632–42.

[16] Cohen J. Statistical Power Analysis for Behavioral Sciences. New York: Academy Press, 1977.

[17] Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–89.

[18] Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191–2.

[19] Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J of Chron Dis 1987;40: 171–8.

[20] Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Controlled Clin Trials 1991;12:142s–58s.

[21] Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.

[22] Beurskens AJHM, de Vet HCW, Köke AJA. Responsiveness of functional status measures in low back pain. Pain 1996;65:71–6.

[23] van der Windt DAWM, van der Heijden GJMG, de Winter A, Koes BW, Devillé W, Bouter L. Validity and responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82–7.

[24] Williams JW, Holleman DR, Simel DL. Measuring shoulder function with the shoulder pain and disability Index. J Rheumatol 1995;22: 727–32.

[25] Heald SL, Riddle DL, Lamb RL. The shoulder pain and disability index: the construct validity and responsiveness of a region-specific disability measure. Phys Ther 1997;77:1079–89.

[26] Guillemin F, Bombardier C, Beaton D. Cross-cutural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417–32.

[27] Deyo RA. Comparative validity of the sickness impact profile and shorter scales for functional assessment for low-back pain. Spine 1986;11:951–4.

[28] Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chaplin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995;20:1899–908.

Appendix

Shoulder disability questionnaire

How to complete this questionnaire: The items of this questionnaire relate to your injured shoulder. If you have trouble with both shoulders, please complete the questionnaire for only one shoulder, that is, the one that was treated (or the side on which you write). When this shoulder hurts, you may experience problems performing daily activities in a normal manner. This list contains 16 statements that shoulder disorder patients have used to describe the situations in which they experience pain and what some of the effects may be. When you read the statements, you may find that some stand out because they apply to your situation today (the past 24 hours). As you go through the list, think of how you felt during the past 24 hours. For each entry, check for yourself whether you performed the mentioned activity.

Examples	NA	Yes	No
1. You did not perform the activity during the past 24 hours, e.g., you *did not* lie on your shoulder during the past 24 hours, put a check mark under NA (not applicable). E.g., My shoulder hurts when I lie on it.	**X**	. .	. .
2. You did perform the activity during the past 24 hours, e.g., you opened or closed a door during the past 24 hours, put a check mark under YES, if your shoulder *did hurt* during the activity. E.g., My shoulder hurts when I open or close a door.	. .	**X**	. .
3. You did perform the activity during the past 24 hours, e.g., you did lean on your elbow or hand during the past 24 hours. If your shoulder *did not hurt* while you were leaning on your elbow or hand, put a check mark under NO. E.g., My shoulder is painful when I lean on my elbow or hand.	. .	. .	**X**

Shoulder Disability Questionnaire, 16 items

	NA	Yes	No
For which shoulder do you complete this questionnaire? Right/Left (circle one).	. .	. .	. .
1. I wake up at night because of my shoulder.	. .	. .	. .
2. My shoulder is hurts when I lie on it.	. .	. .	. .
3. Because of my shoulder I have trouble putting on a coat or a sweater.	. .	. .	. .
4. My shoulder hurts during my usual daily activities.	. .	. .	. .
5. My shoulder hurts when I move my arm.	. .	. .	. .
6. My shoulder hurts when I lean on my elbow or hand.	. .	. .	. .
7. My shoulder hurts when I write or type.	. .	. .	. .
8. My shoulder hurts when I hold my car steering wheel or my bike handlebars.	. .	. .	. .
9. My shoulder hurts when I lift and carry something.	. .	. .	. .
10. My shoulder hurts when I reach or grasp above shoulder level.	. .	. .	. .
11. My shoulder hurts when I open or close a door.	. .	. .	. .
12. My shoulder hurts when I bring my hand towards my buttocks.	. .	. .	. .
13. My shoulder hurts when I bring my hand towards my lower back.	. .	. .	. .
14. My shoulder hurts when I bring my hand towards the back of my head.	. .	. .	. .
15. I rub my shoulder more than once during the day.	. .	. .	. .
16. I am irritable and bad tempered with people because my shoulder hurts.	. .	. .	. .